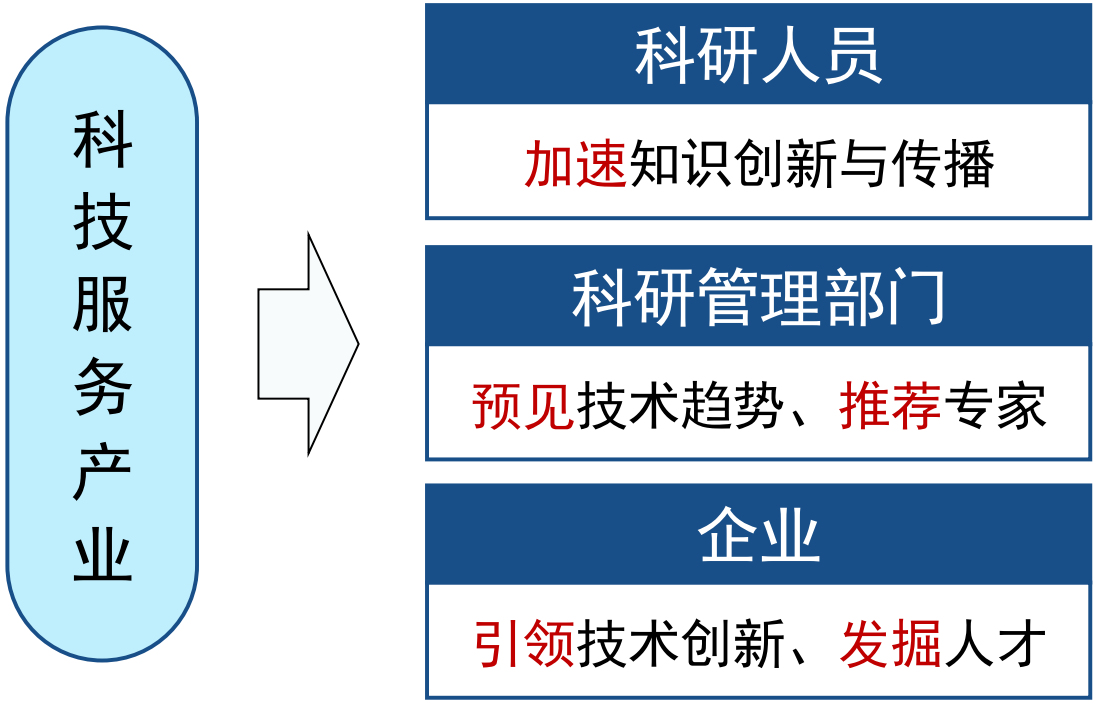
知识驱动的科技情报挖掘

Jie Tang
Tsinghua University

[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. KDD'08. pp.990-998.

科技情报

- 科技情报是从海量科技数据中挖掘出来的知识创新和传播规律
- 帮助科研人员、国家科研管理机构和企业提升科技生产力

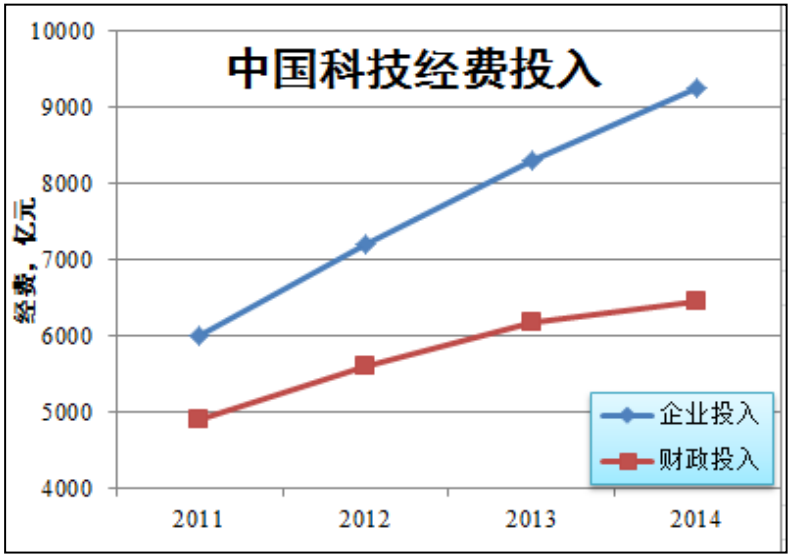


情报驱动的科技服务产业

- 谁掌握了科技发展规律，谁就掌握了经济发展引擎！

❑国民经济

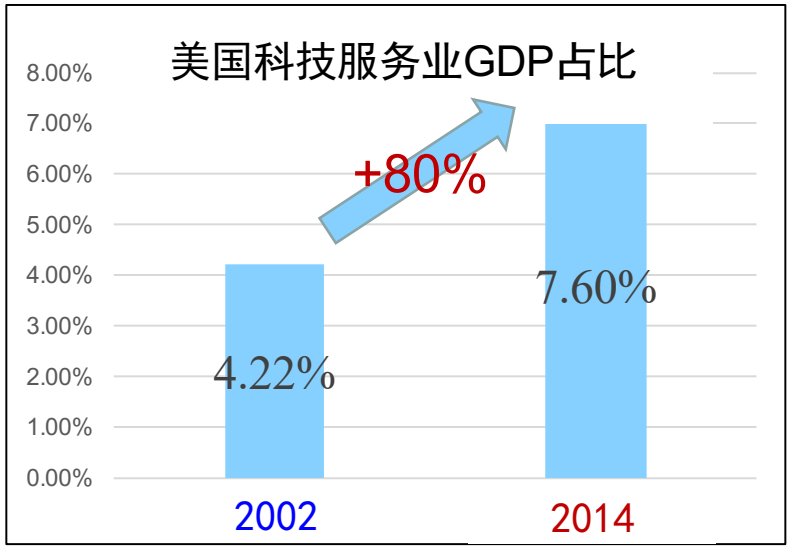
到2020年，科技服务业产业规模有望达到8万亿元



国家统计局，2014

❑对比美国科技经济

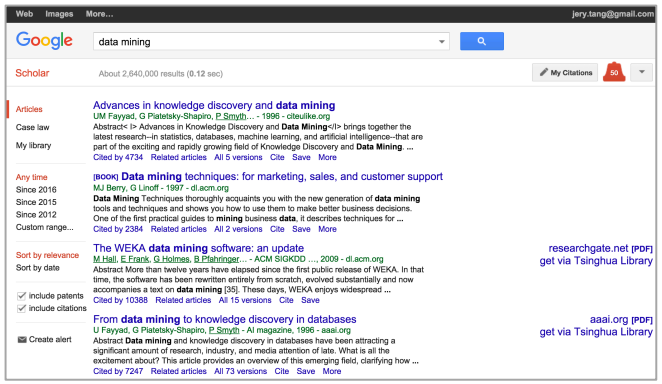
美国科技服务业GDP占比在过去几年增长快速

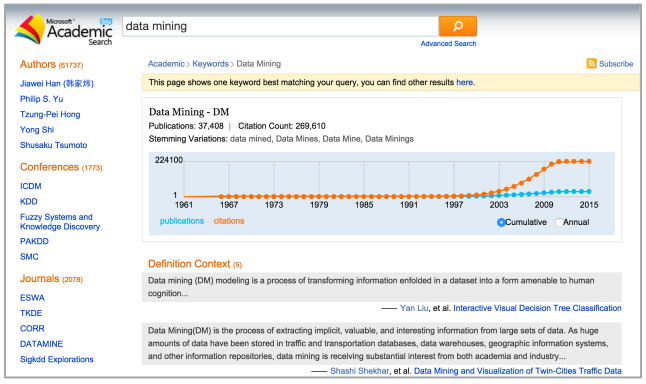


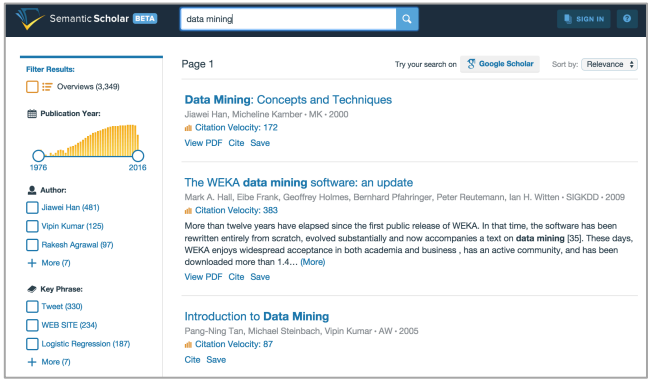
求是网，2015

Background

公司	系统名称	发布时间	特性描述
谷歌	Google Scholar	2004	大规模学术论文索引与快速检索
微软	Academic Search	2014	学术搜索
汤森路透	Web of Science	2008	科学引用指标、交叉学科引用指标
爱思唯尔	Scopus	2004	全学科、交叉学科文献收集







Background

传统情报分析

以文献为中心的关键词情报分析

难以满足科技服务产业快速发展的需求

语义挖掘

智能型情报挖掘

构建知识图谱

隐含关联

挖掘人才网络

形成知识驱动的情报挖掘技术体系

科技情报挖掘面临的新挑战

1 情报信息以碎片化的形式分散在异构、多源数据中，知识获取难度大；

美国NSF和Army（仅2015年就超过8千万美元）- Science of Science

2 情报网络结构复杂、交互行为动态多样，挖掘情报隐含关联关系十分困难；

欧盟第8框架H2020（6项关于科技情报分析，共21项学科领域项目）

今年截止目前为止的Nature和Science杂志中已有23篇关于科技大数据挖掘的论文

3 情报网络规模大、数据多维，匹配效率低是智能型情报的计算瓶颈。

多个顶尖学术研究机构（MIT，哈佛，康奈尔，芝加哥等）建立相关团队

AMiner.cn： 知识驱动的科技情报挖掘平台

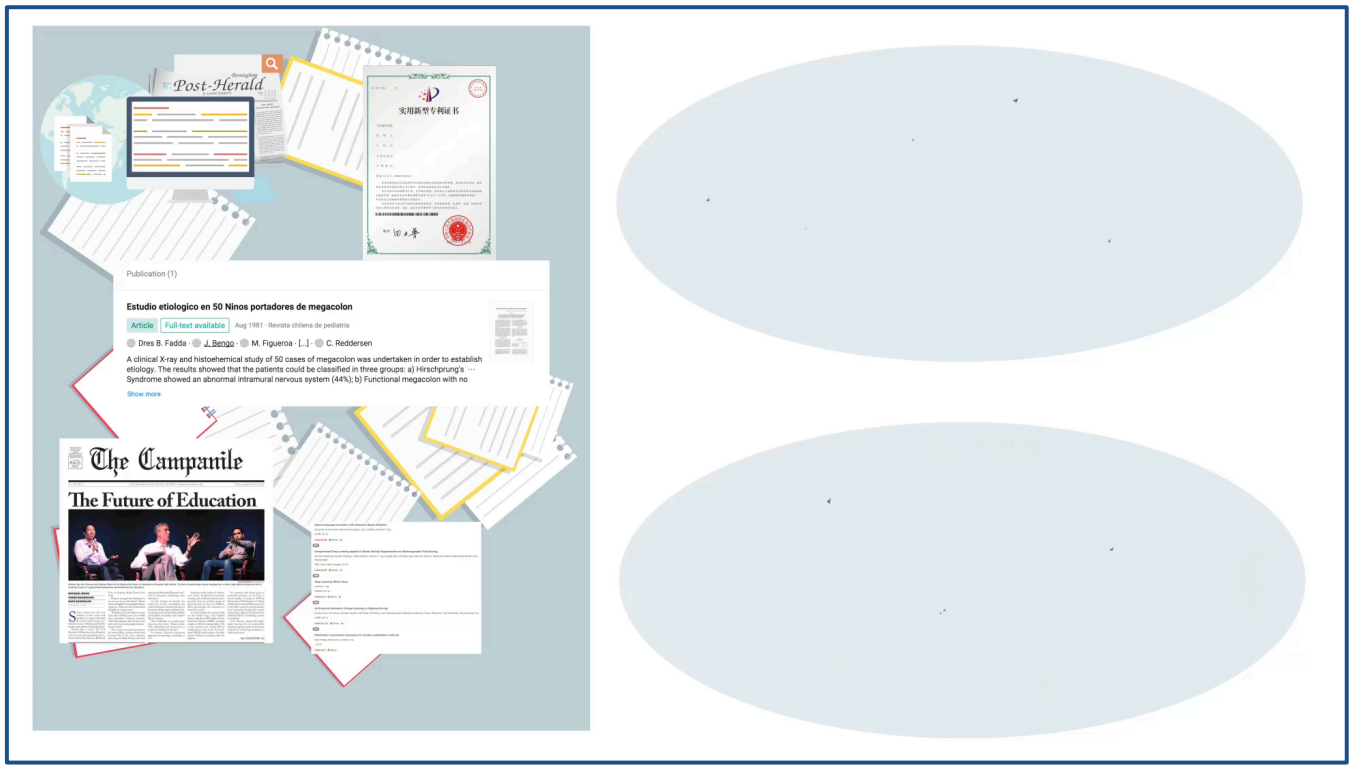
知识获取
科技知识图谱构建

关联
挖掘

隐含关联
知识到情报挖掘

快速
算法

语义匹配
情报到智能服务



AMiner EN Dashboard | Login

AMiner.cn

MINING DEEP KNOWLEDGE FROM SCIENTIFIC NETWORKS

Whatever comes to your mind

Advanced

Hot Topics

Data Mining

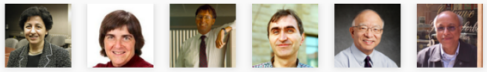
Social Network

Machine Learning

Deep Learning

Computer Vision

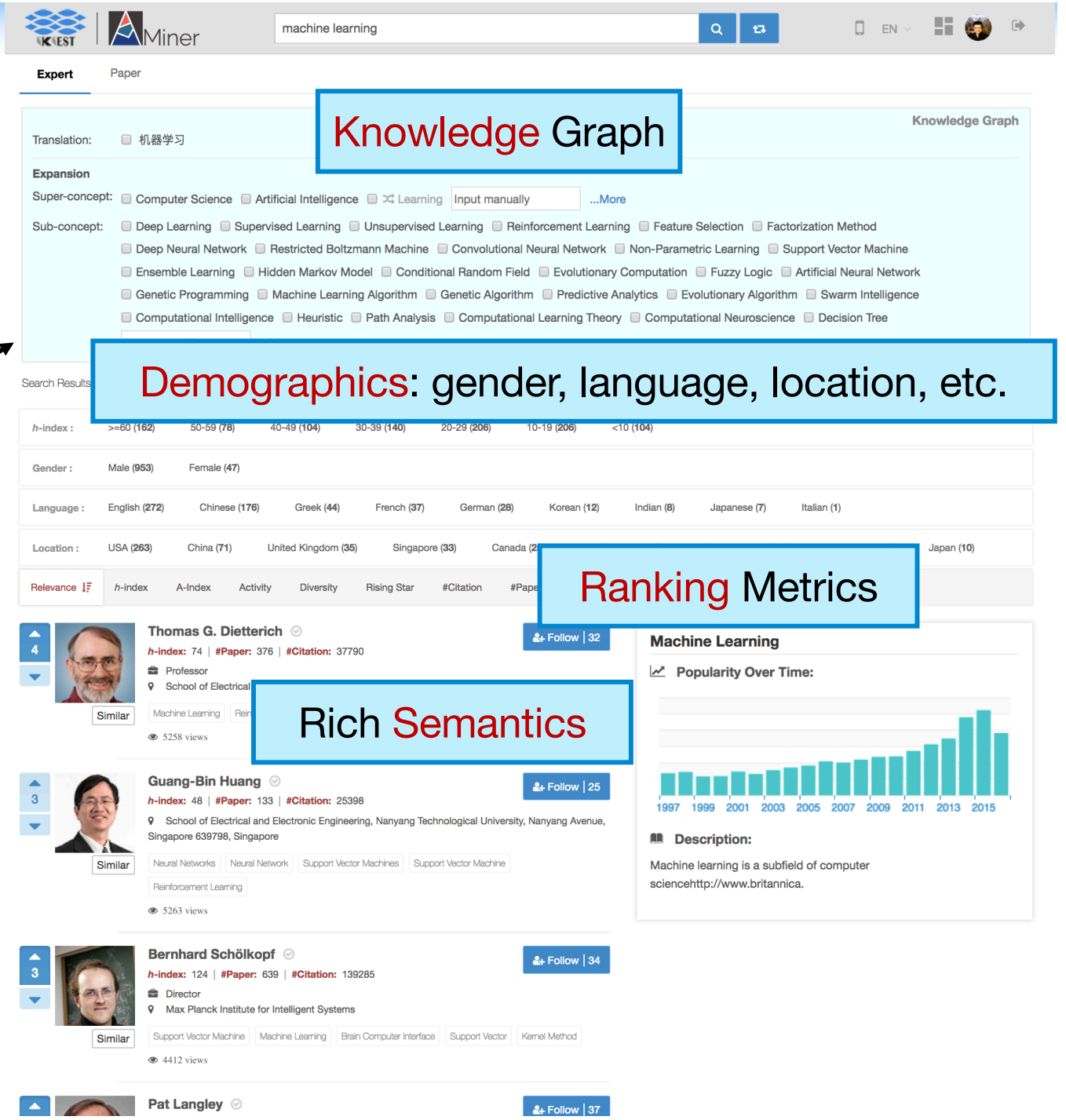
Database



136,722,706	101,390,721	7,854,301	133,196,029
RESEARCHERS	PUBLICATIONS	CONCEPTS	CITATIONS

Open Services
—语义搜索

机器学习
专家



https://aminer.cn

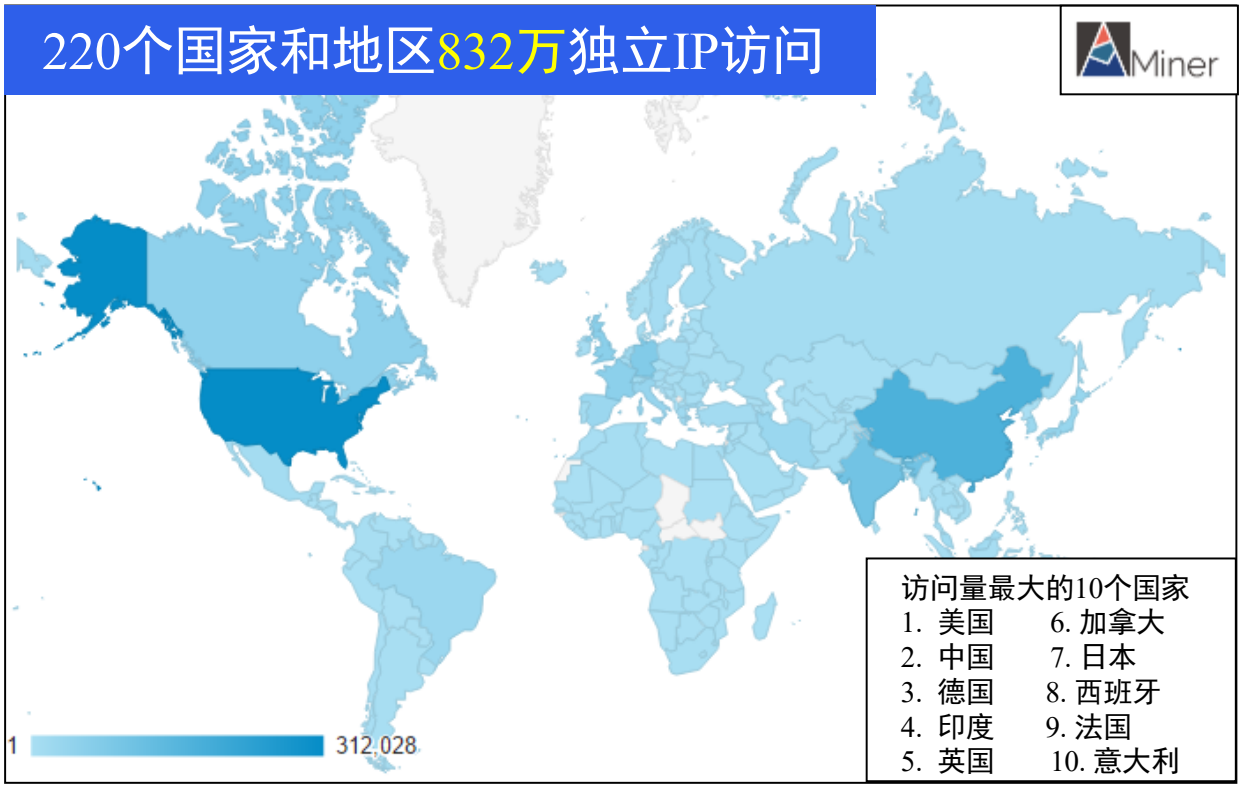
Open Report—技术发展报告

https://www.aminer.cn/research_report/articlelist

AMiner访问量及用户分布

Google Analytics

220个国家和地区832万独立IP访问



访问量最大的10个国家

1. 美国
2. 中国
3. 德国
4. 印度
5. 英国
6. 加拿大
7. 日本
8. 西班牙
9. 法国
10. 意大利

爱尔兰研究机构
DERI资深研究员
P. Buitelaar

Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food

Fergal Monaghan, Georgeta Bordea, Krystian Samp, and Paul Buitelaar

fron cannot be positioned relative to all of these here given the limited space, we now compare Saffron with the most related work: ArnetMiner[8] [6], a well-known state of the art "academic researcher social network search" tool.

ArnetMiner has an emphasis on clasification and consists of two main parts. In the first part, probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1] are extended and a unified topic model for papers, authors and conferences is proposed. It seems only the content of the papers is analysed, and structured data, such as social connections, is not considered. They cluster all

AMiner is a well-known state of the art tool……

在线运行超过十年

科研数据下载230万次，年均数据访问量超过1100万次

核心技术

—How to populate a semantic-based profile database for researchers?

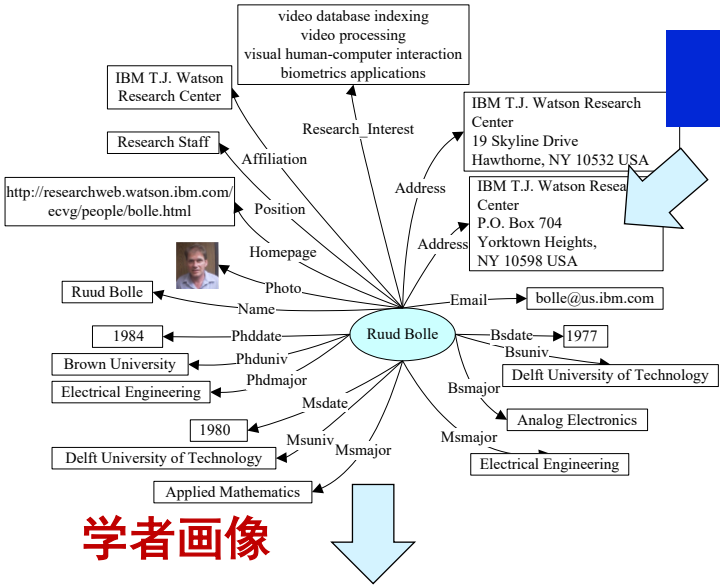
Edward Feigenbaum
专家系统之父
图灵奖

Tim Berners Lee
WWW创始人
图灵奖

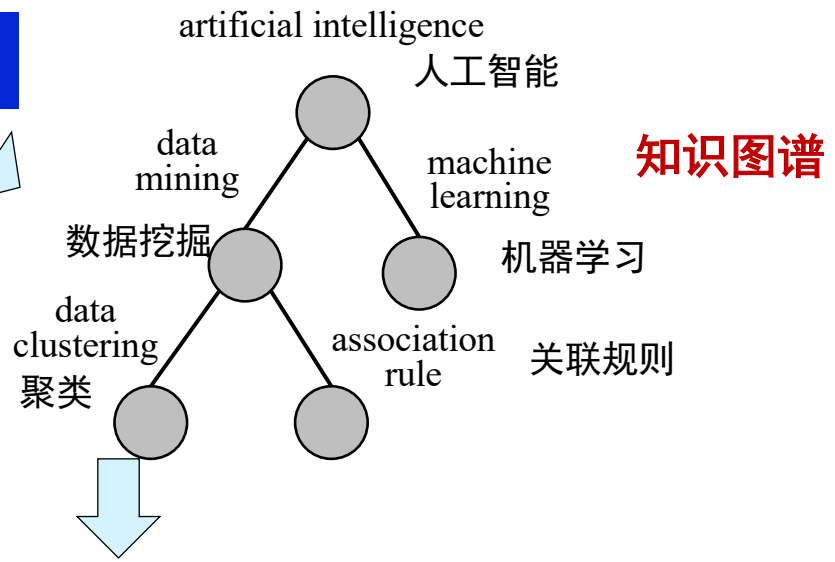
大数据 ⇨ 知识 ⇨ 智能

*人工智能的两个重要阶段：大规模知识库 + 智能服务

Architecture—以AI为例



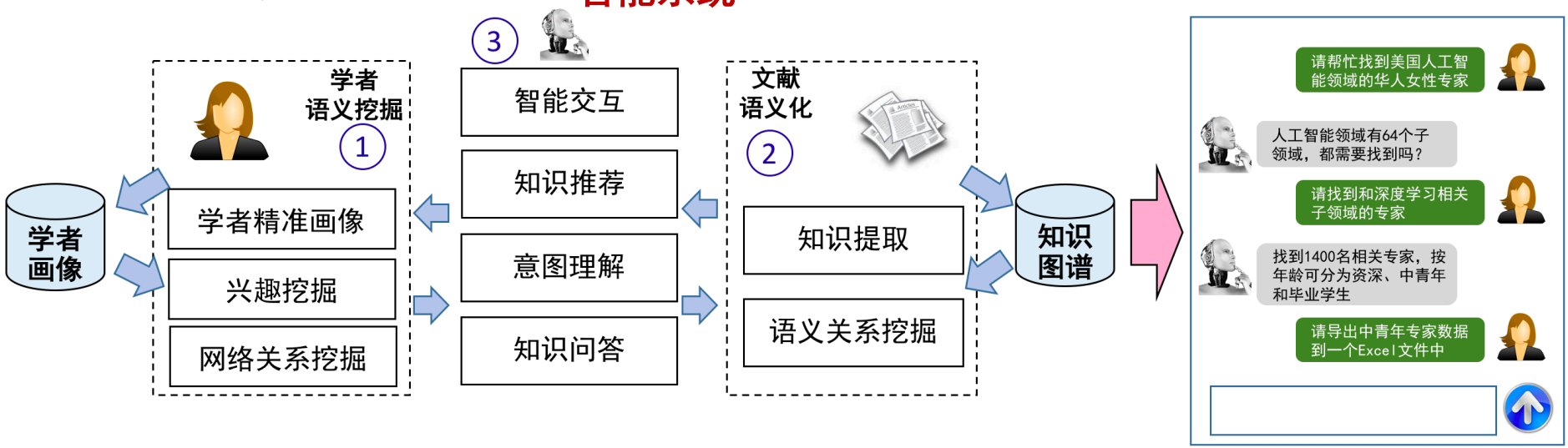
人工智能



知识图谱

学者画像

智能系统



Technology Overview

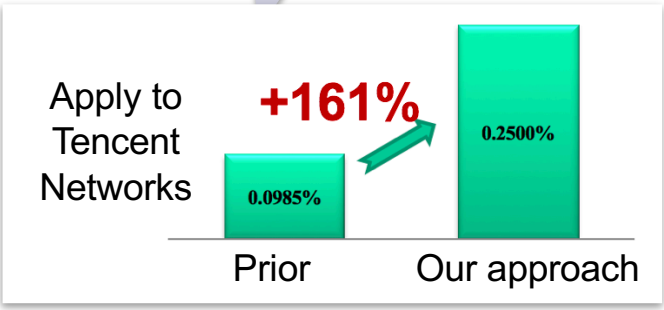
Data->Semantics
Search->Intelligence

智能服务

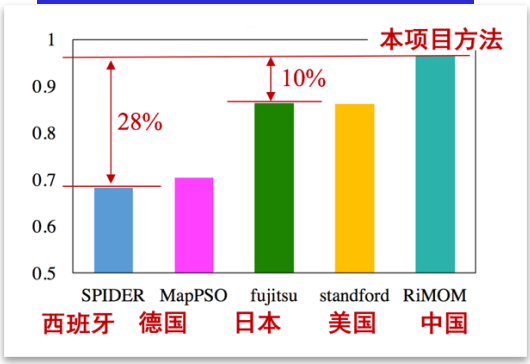
语义集成

数据语义化

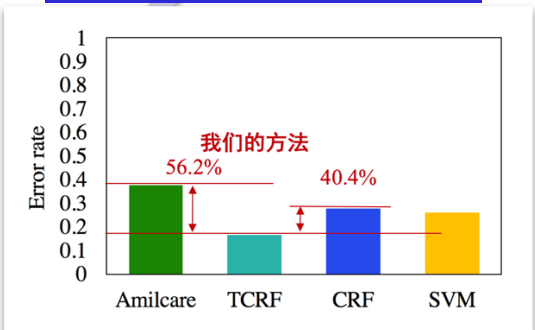
Recommendation accuracy +161%



15 champions in the past 7 years



Error rate reduced 40-56%

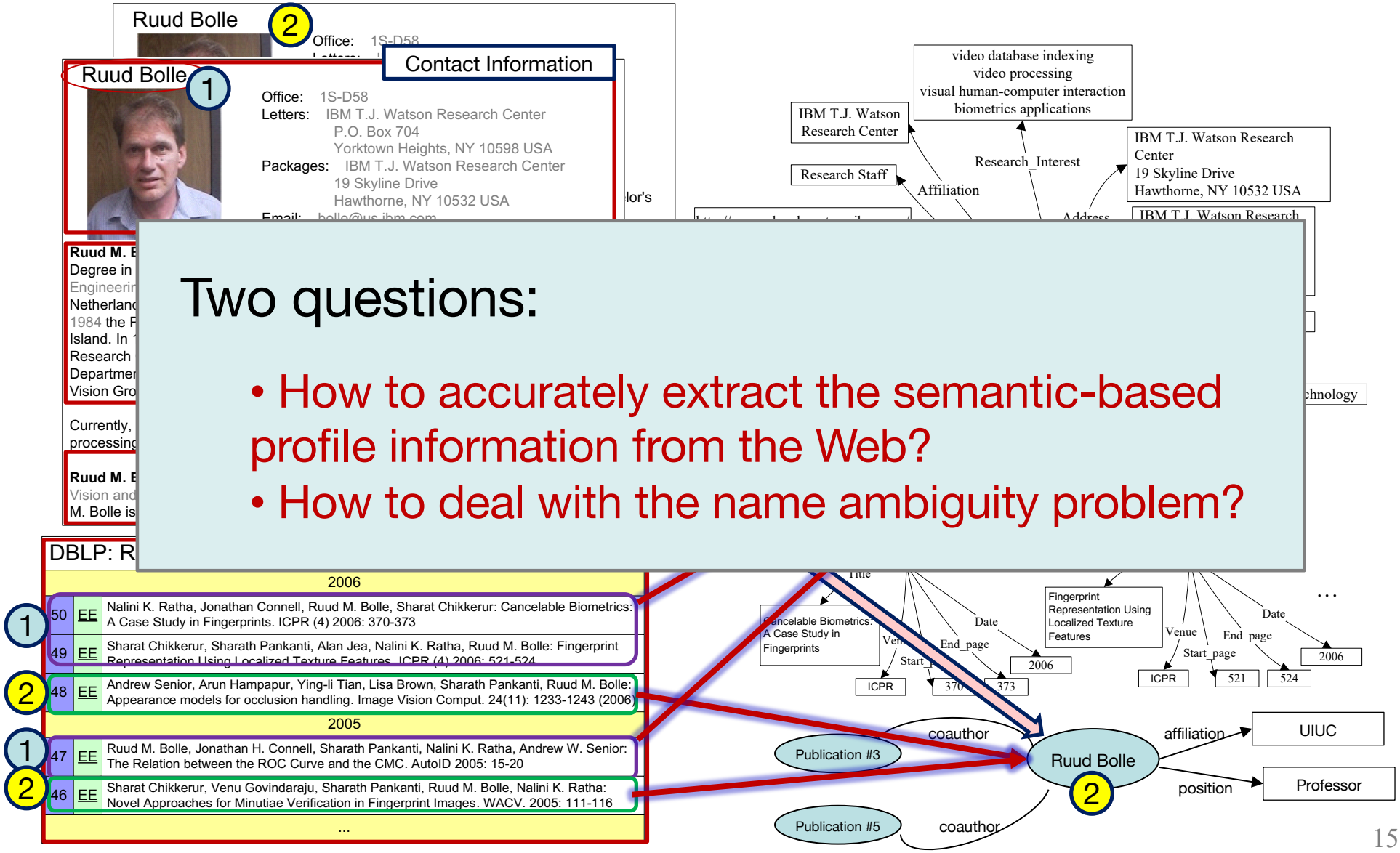


Reported by UN

#citation>10,000, published ~200 papers on major journals and conferences

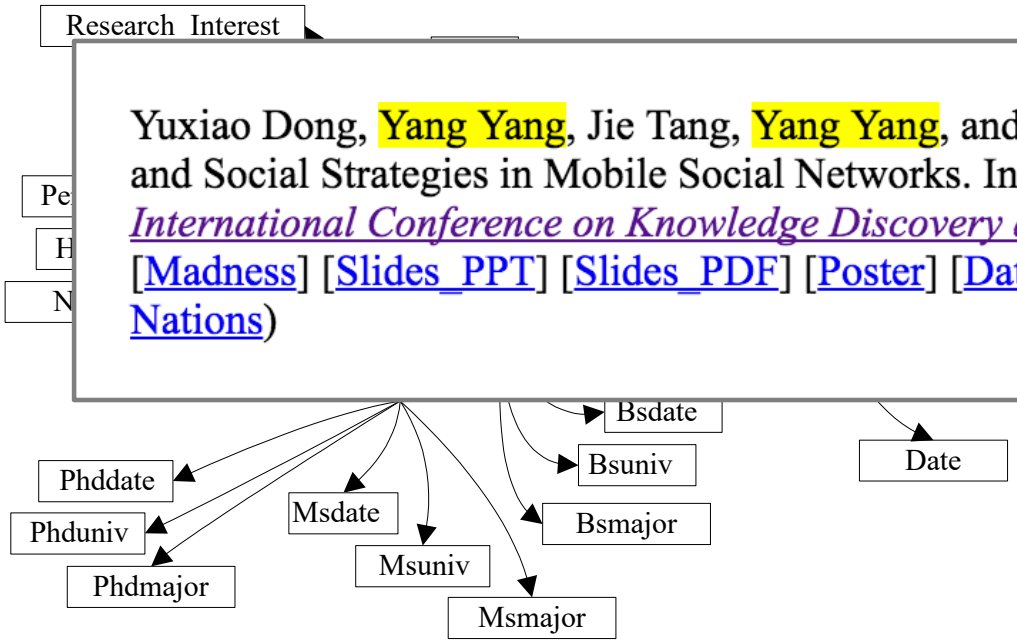
Extracting Profile Semantics from the Web

(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)



Researcher Profile Extraction[1,2]

70.60% of the researchers have at least one homepage or an introducing page



There are a large number of person names having the ambiguity problem

Even 3 "Yi Li" graduated the author's lab

70% moved at least one time

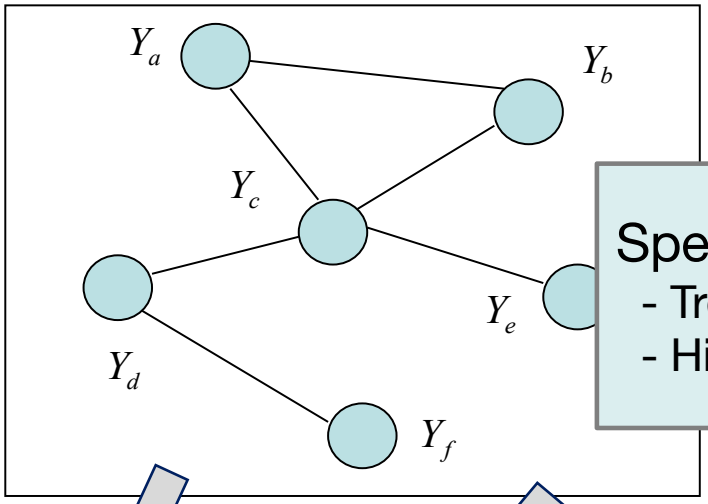
[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. KDD'08. pp.990-998.

Our Approach Picture

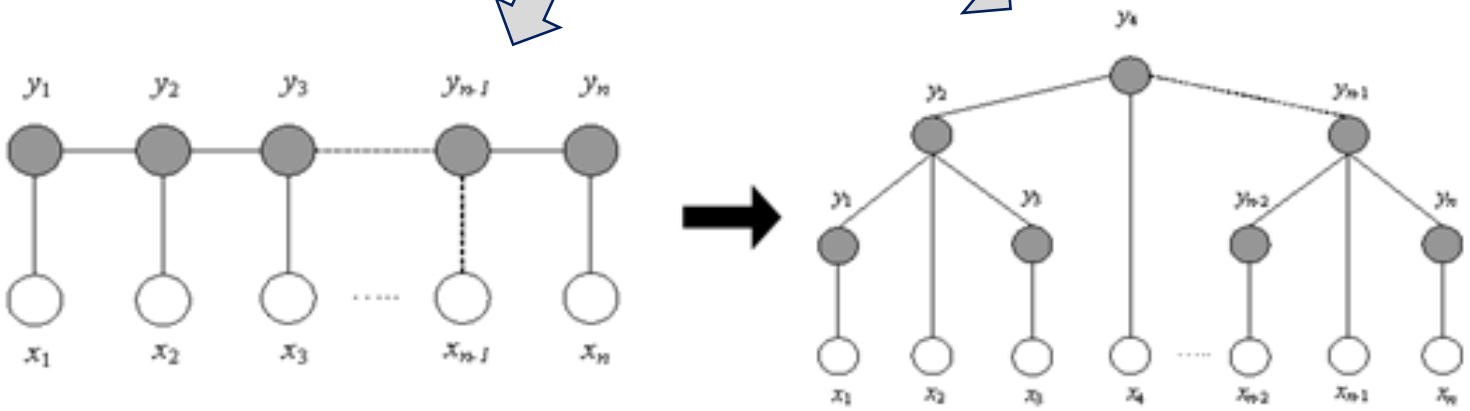
Markov Property:

$$P(Y_i \mid \{Y_j \mid Y_j \neq Y_i\})$$

$$= P(Y_i \mid \{Y_j \mid Y_j \sim Y_i\})$$

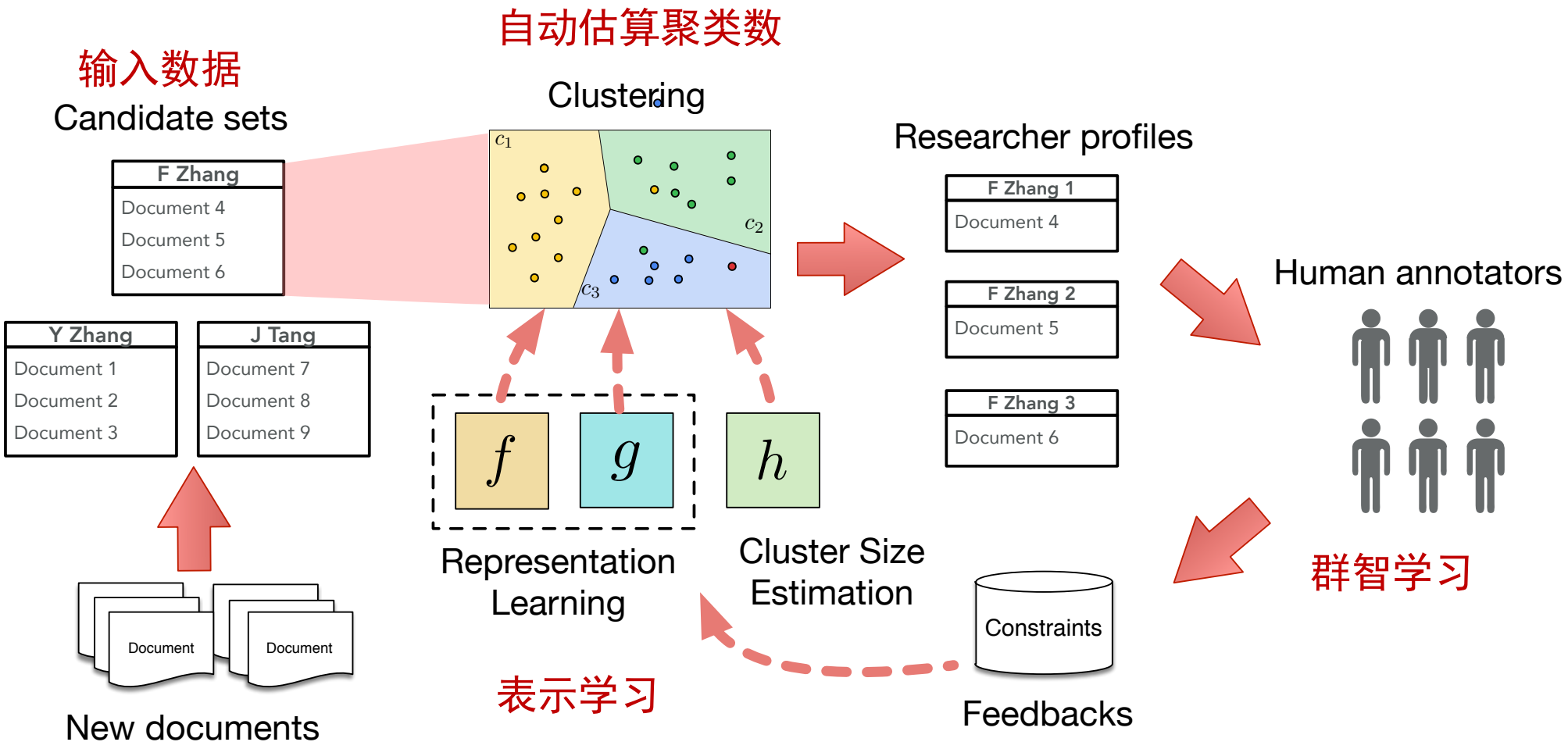


Special Cases:
- Tree Conditional Random Fields
- Hidden Markov Random Fields



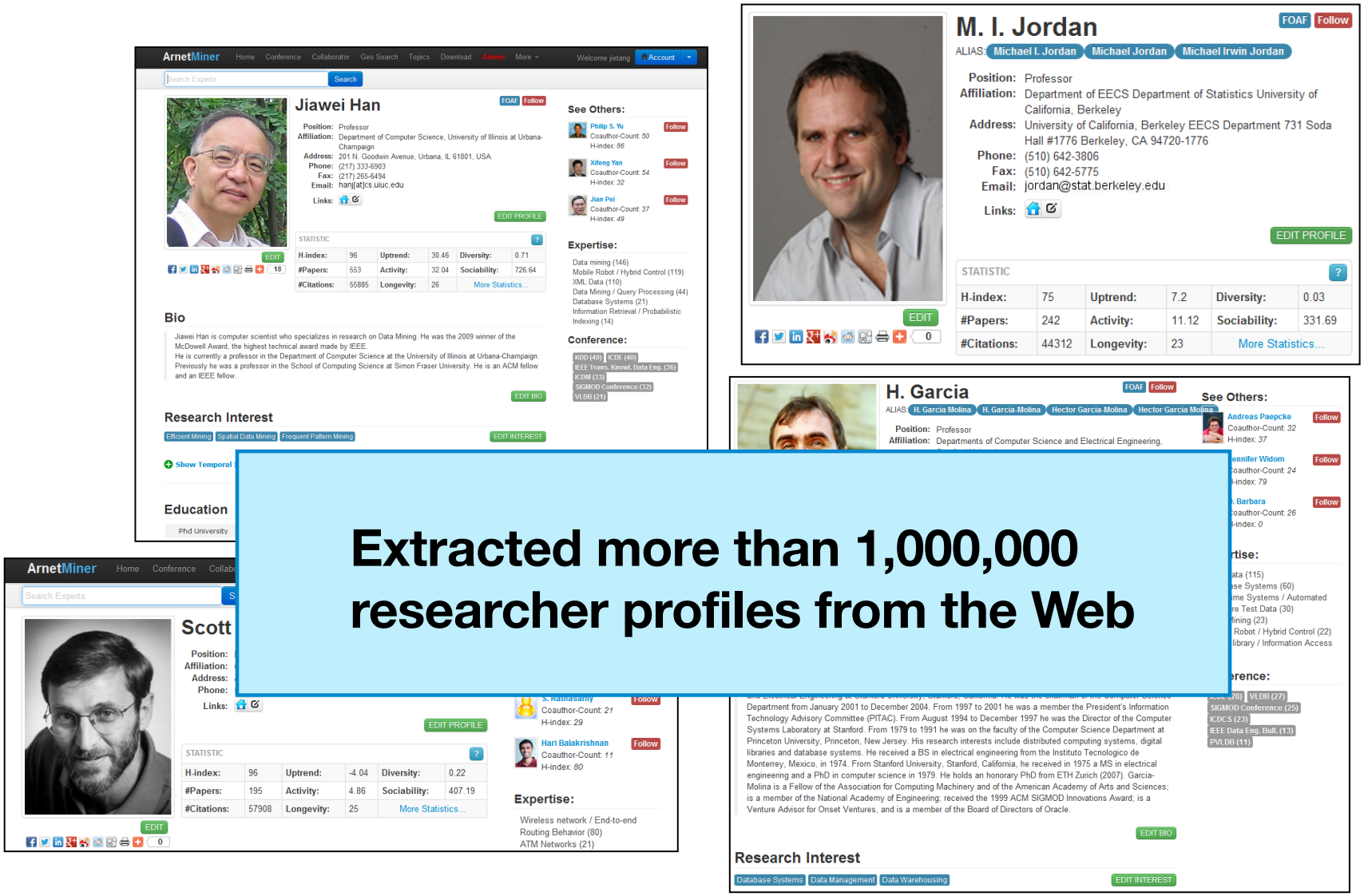
$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{e \in \{E^{pc}, E^{cp}, E^{ss}\}, j} \lambda_j t_j(e, y|_e, x) + \sum_{v \in V, k} \mu_k s_k(v, y|_v, x) \right)$$

基于深度学习的在线数据集成框架



[1] Y. Zhang, F. Zhang, P. Yao, and J. Tang. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. KDD'18. pp. 1002-1011.

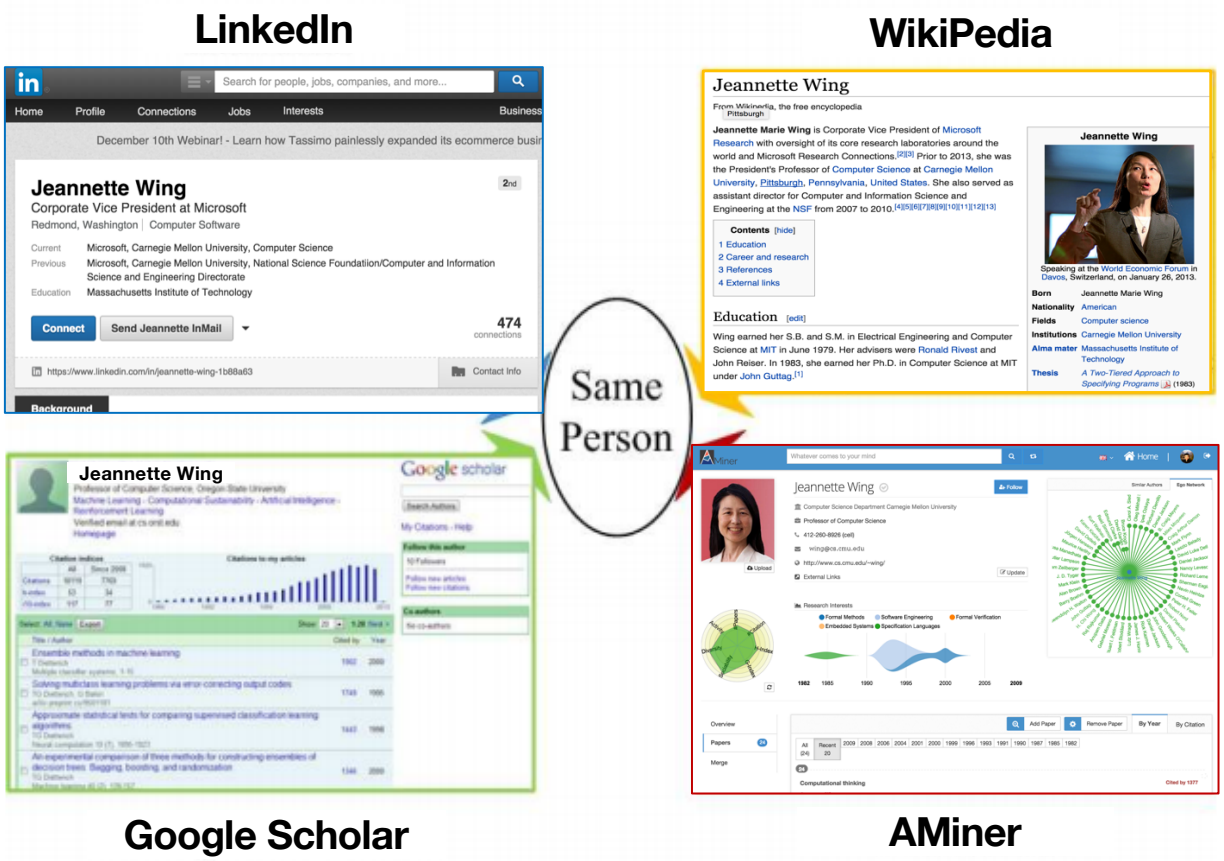
Researcher Profile Database



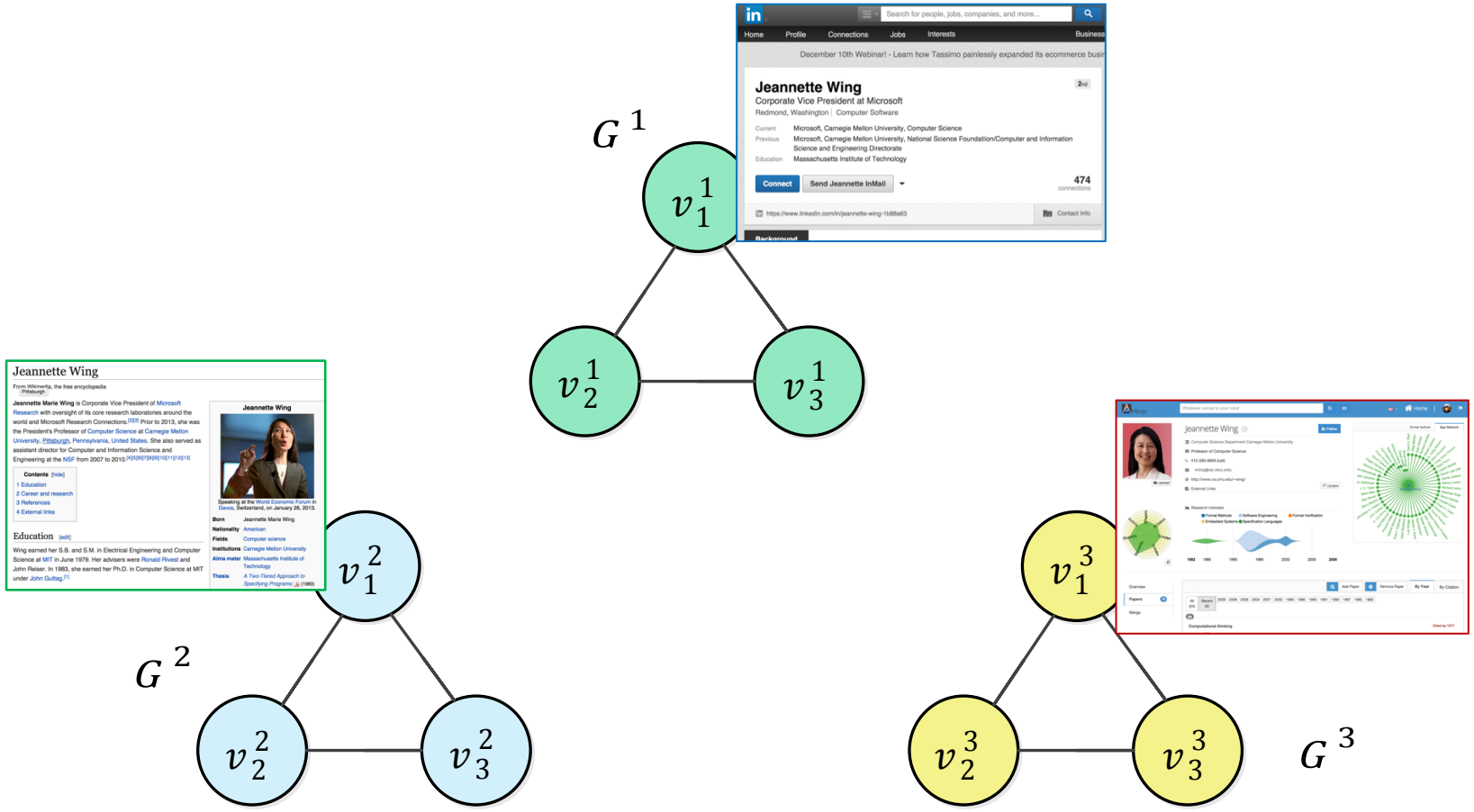
[1] J. Tang, L. Yao, D. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.

Linking Semantics across Networks

- Identifying users from multiple heterogeneous networks and integrating semantics from the different networks together.

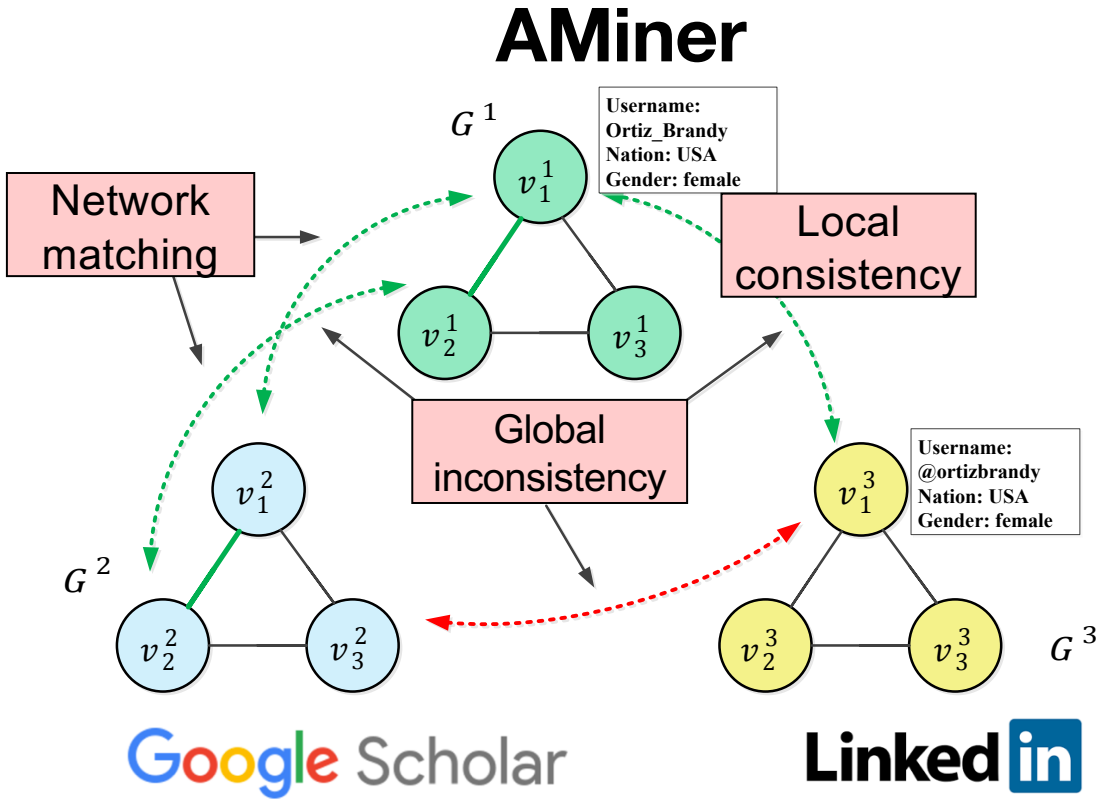


Considering the networks…



Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency



DEFINITION 2 (GLOBAL INCONSISTENCY). *Given a set of social networks* $\mathbf{G}$*, a set of user pairs* X *and the corresponding labels* Y*, if there exists a sequence of user pairs* $\langle \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_n} \rangle$*, such that*

$$\forall i = i_1, i_2, \cdots, i_n, y_i = 1$$

and

$$\forall k = 1, 2, \cdots, n-1, \mathcal{V}^2_{i_k} = \mathcal{V}^1_{i_{k+1}}$$

and

For the pair $\langle \mathcal{V}^2_{i_n}, \mathcal{V}^1_1 \rangle$*, if the corresponding label* $y_j = 0$

then we say that the assigned labels Y *causes global inconsistency given* $\mathbf{G}$ *and* X*.*

Avoid "global inconsistency"

COSNET: Connecting Social Networks with Local and Global Consistency

- Input: $\mathbf{G}=\{G^1, G^2, \ldots, G^m\}$, with $G^k=(V^k, E^k, R^k)$
- Formalization: $\mathbf{X}=\{x_i\}$, all possible pairwise matchings and each corresponds to $y_i \in \{1,0\}$

- COSNET: an energy-based model

$$Y^* = \arg\max E(Y, X)$$

[1] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. KDD'15, page 1485-1494.